

Clinical Data Classification with Quantum Kernel Alignment

QTM 2026 – Extended Abstract

Florian Heininger ¹, Jeanette Miriam Lorenz ^{3,4}, Matthias Weidemüller ², and Niels Halama ¹

¹Hi-TRON, German Cancer Research Center, Mainz, Germany

²Physics Institute, Heidelberg University, Heidelberg, Germany

³Fraunhofer IKS, Munich, Germany

⁴Faculty of Physics, LMU Munich, Munich, Germany

Abstract

Quantum kernel methods are among the most promising near-term applications of quantum machine learning, with the quantum feature map controlling a model’s ability to capture meaningful correlations from little labelled data. We present a systematic approach that aligns a quantum-inspired feature map to highly imbalanced and scarce clinical datasets through kernel–target alignment (KTA) optimization, and feeds the resulting Gram matrix into a classical support vector machine (SVM). On six clinical datasets of therapeutic and prognostic relevance that are effectively intractable for a classical radial basis function (RBF) kernel SVM, the aligned quantum kernel improves the AUC by ≈ 0.22 on average, while matching the classical baseline on a balanced benchmark. This reveals a practical NISQ-era pathway to analyze complex clinical data that is otherwise resistant to conventional machine learning.

1 Introduction

Quantum machine learning (QML) has emerged at the intersection of quantum computing and machine learning [1]. Current devices are limited by noise and short coherence times – the noisy intermediate-scale quantum (NISQ) regime – yet among the most promising NISQ applications are quantum kernel methods (QKMs), particularly for their potential to reduce the amount of required training data [2]. Here, data-embedding circuits, the *quantum feature maps*, encode classical data into quantum states and can enhance the expressivity of the induced kernel. High expressivity is especially valuable for rare clinical diagnoses, where heavily imbalanced and scarce datasets arise naturally: in personalized oncology, small and heterogeneous datasets result from rare disease entities (e.g. pancreatic cancer) and small patient subgroups. Combining small sample sizes with high feature dimension – the curse of dimensionality [3] – such datasets hinder robust model development and form a fundamental barrier to clinical progress. Because most QML studies use balanced, strongly correlated benchmark data, their effect on real clinical data remains unclear; recent work even suggests the intrinsic data complexity sets a threshold for a possible quantum advantage [4]. We address this gap directly with clinically relevant, strongly imbalanced datasets.

2 Methods

Quantum embedding kernels. A kernel based SVM separates classes with a maximum-margin hyperplane in a feature space implicitly defined by a feature map ϕ , using only inner products $\kappa(x_i, x_j) = |\langle \phi(x_j), \phi(x_i) \rangle|^2$; the classical baseline is the Gaussian RBF kernel $\kappa_{\text{RBF}}(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$. A quantum feature map instead encodes an m -dimensional point x into an n -qubit state $|\psi(x)\rangle = U(x) |0^{\otimes n}\rangle$, and the kernel entry is estimated as the overlap $|\langle 0^{\otimes n} | U^\dagger(x_j) U(x_i) | 0^{\otimes n} \rangle|^2$, approximated by the fraction of all-zero outcomes over the measurement

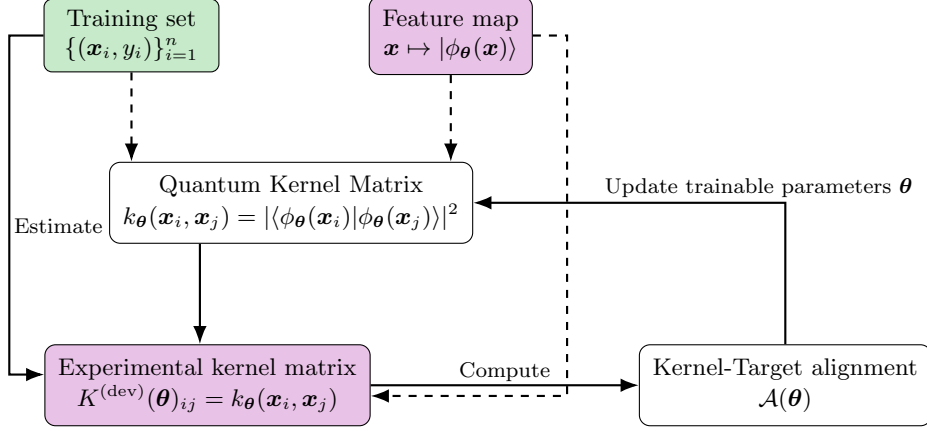


Figure 1: Quantum kernel alignment procedure

shots. We use a Hamiltonian evolution feature map developed for many-body problems and proven powerful in a QKM setting [5, 6],

$$|x_i\rangle = \left(\prod_{j=1}^n e^{-i \frac{t}{T} x_{ij} H_j^{XYZ}} \right)^T U_{\text{ent}}^L |0^{\otimes(n+1)}\rangle, \quad H_j^{XYZ} = X_j X_{j+1} + Y_j Y_{j+1} + Z_j Z_{j+1}, \quad (1)$$

where X_j, Y_j, Z_j are Pauli operators, t the evolution time, T the number of Trotter steps, and U_{ent}^L applies L strongly entangling layers of single-qubit rotations and entanglers [7]. The map uses $n+1$ qubits to embed n features.

Kernel–target alignment. The choice of feature map dominates QKM performance, so we tune the parameters $\theta = (t, T, L)$ to the data via the kernel–target alignment, which measures the agreement between the kernel matrix K and an ideal label-derived target $K_T = yy^\top$ [8],

$$\mathcal{A}(K, K_T) = \frac{\langle K, K_T \rangle_F}{\sqrt{\langle K, K \rangle_F \langle K_T, K_T \rangle_F}} = \frac{\langle K, K_T \rangle_F}{\sqrt{\text{Tr}(K^2) N}}, \quad \mathcal{A} \in [-1, 1], \quad (2)$$

with N training samples. On a stratified subsample $(X_s, \mathbf{y}_s)$ of the training set we maximize the centered alignment $C(\theta) = \mathcal{A}(K_\theta(X_s, X_s), T_{\text{target}}(\mathbf{y}_s))$, where $T_{\text{target}}(\mathbf{y}_s) = \mathbf{y}_s \mathbf{y}_s^\top$ is the label-derived target K_T of Eq. (2) formed from the subsample labels: the continuous time t is updated by a momentum-based projected gradient ascent (Euclidean projection onto $[0, t_{\text{max}}]$), while the discrete T and L are refined by a local hill-climb over nearest neighbors. Each candidate $\theta^{(k+1)}$ is accepted or rejected by comparing $C(\theta^{(k+1)})$ with $C(\theta^{(k)})$, and the best $\hat{\theta} = \arg \max_k C(\theta^{(k)})$ is retained (Fig. 1). The resulting Gram matrix is inserted into a classical SVM whose remaining hyperparameters are fixed by a random search followed by an AUC-scored grid search. As accuracy is misleading under strong imbalance, we evaluate models with the area under the ROC curve (AUC) [9].

3 Datasets and Results

We study seven datasets of clinical relevance spanning a range of imbalance and sample size (Table 1). *Breast Cancer Wisconsin* is a near-balanced public benchmark [10] and *Stroke Prediction* a public, strongly imbalanced one. The remaining sets are synthetic, anonymized derivatives of routine clinical data from maximum-care cancer centers [11]: *PDAC* (rare lethal pancreatic cancer), *Signalling Molecules* (molecular-tumor-board target identification), *CRC* (heterogeneous metastatic colorectal cancer), *Solid Cancers* (pediatric immunotherapy response), and *Cold MSS* (rare responders within checkpoint-inhibitor-resistant “cold” MSS gastrointestinal tumors).

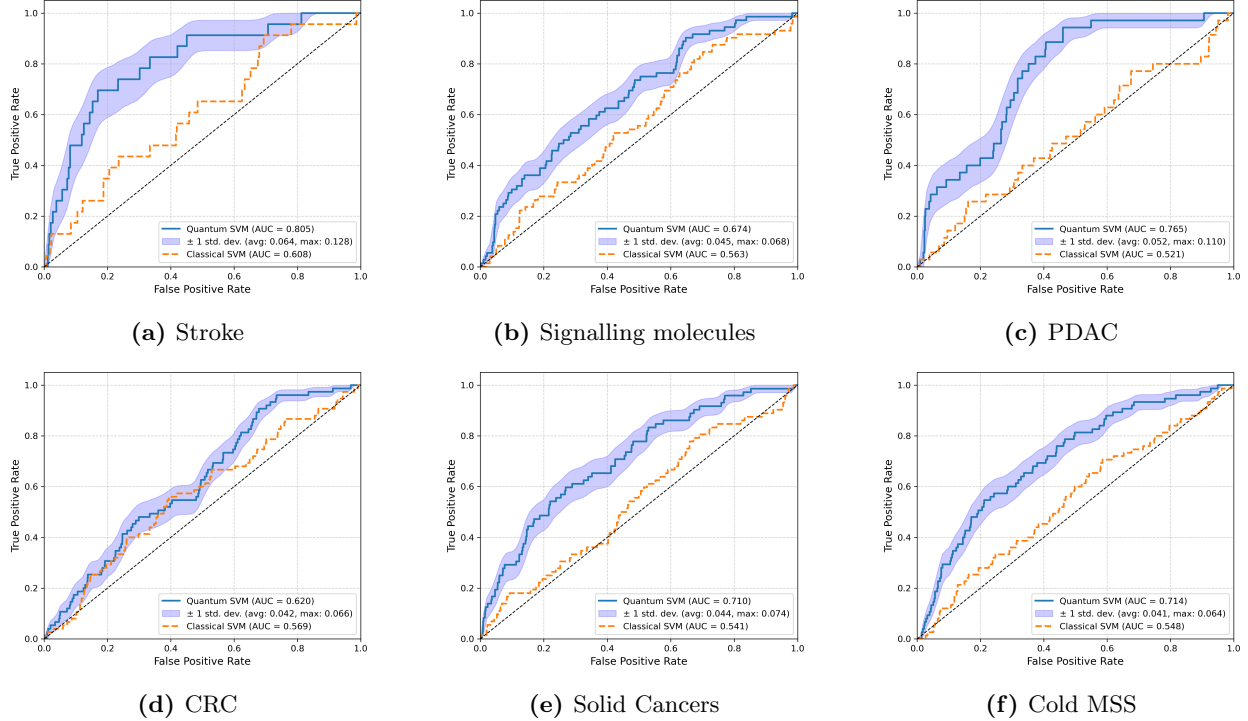


Figure 2: ROC curves of the aligned quantum kernel SVM (q-inspired) compared with a classical RBF kernel SVM across the six imbalanced clinical datasets; the aligned quantum kernel consistently achieves higher true-positive rates on the minority class. AUC values are listed in Table 2.

On the balanced breast-cancer benchmark the aligned quantum kernel matches the classical RBF kernel (AUC 0.99 vs. 0.99), confirming that alignment recovers the strong-baseline regime. On the heavily imbalanced data it consistently outperforms the RBF SVM (Table 2, Fig. 2): for the stroke dataset reduced to 30% of its samples by stratified group sampling, AUC rises from 0.61 to 0.81 (and from 0.67 to 0.73 on the full set), and the oncology datasets improve by up to 0.25 AUC (PDAC 0.52 $\rightarrow$ 0.77, Solid Cancers 0.54 $\rightarrow$ 0.71, Cold MSS 0.55 $\rightarrow$ 0.71). Averaged over the six imbalanced datasets, the systematic feature-map alignment yields a mean AUC gain of ≈ 0.22 .

4 Discussion

We demonstrate a practical pathway to approach imbalanced and scarce clinical datasets with QKMs by systematically aligning quantum feature-map parameters via kernel-target alignment. Such clinical and trial datasets – small, noisy, and strongly imbalanced – are notoriously hard to analyze, yet their stratification is directly tied to therapy choice, prognosis, and the identification of responding patient subgroups. The advantage grows with the difficulty of the data: the more size-limited and feature-heterogeneous the dataset, the more clearly the quantum-inspired kernel

Table 1: Overview of the datasets: Sample-size, level of imbalance and number of features.

Dataset	Size	Imbalance	Feat.
Breast Cancer	569	1.7:1	30
Stroke Prediction	5110	19.5:1	10
Signalling Molecules	3806	14.8:1	9
PDAC	5096	20.4:1	10
CRC	4966	18.9:1	10
Solid Cancers	3808	14.8:1	9
Cold MSS	5108	19.5:1	10

Table 2: AUC comparison of a classical RBF kernel vs. the aligned quantum (q-inspired) kernel.

Dataset	AUC (RBF)	AUC (q-insp.)
Breast Cancer	0.99	0.99 $\pm$ 0.01
Stroke Prediction	0.61	0.81 $\pm$ 0.06
Signalling Molecules	0.56	0.67 $\pm$ 0.05
PDAC	0.52	0.77 $\pm$ 0.05
CRC	0.57	0.62 $\pm$ 0.04
Solid Cancers	0.54	0.71 $\pm$ 0.04
Cold MSS	0.55	0.71 $\pm$ 0.04

outperforms its classical counterpart, while still matching it on easy, balanced data. The method is not specific to oncology, and as NISQ hardware matures, more expressive feature maps define a hardware-scaling landscape for this concrete near-term application of quantum computing.

References

- [1] Richard P. Feynman. Simulating physics with computers. *International Journal of Theoretical Physics*, 21(6):467–488, 1982.
- [2] M. Emre Sahin, Benjamin C. B. Symons, Pushpak Pati, Fayyaz Minhas, Declan Millar, Maria Gabrani, Stefano Mensa, and Jan Lukas Robertus. Efficient Parameter Optimisation for Quantum Kernel Alignment: A Sub-sampling Approach in Variational Training. *Quantum*, 8:1502, October 2024.
- [3] Visar Berisha, Chelsea Krantsevich, P. Richard Hahn, Shira Hahn, Gautam Dasarathy, Pavan Turaga, and Julie Liss. Digital medicine and the curse of dimensionality. *npj Digital Medicine*, 4(1):153, 2021.
- [4] Christophe Pere. Data complexity: a threshold between classical and quantum machine learning – part i, 2025.
- [5] Sebastian Egginger, Alona Sakhnenko, and Jeanette Miriam Lorenz. A hyperparameter study for quantum kernel methods. *Quantum Machine Intelligence*, 6(2):44, 2024.
- [6] Brian Coyle, Daniel Mills, Vincent Danos, and Elham Kashefi. The born supremacy: quantum advantage and training of an ising born machine. *npj Quantum Information*, 6(1):60, 2020.
- [7] Maria Schuld, Alex Bocharov, Krysta M. Svore, and Nathan Wiebe. Circuit-centric quantum classifiers. *Physical Review A*, 101(3), March 2020.
- [8] Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz Kandola. On kernel-target alignment. In *Proceedings of the 15th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS’01, page 367–373, Cambridge, MA, USA, 2001. MIT Press.
- [9] J. Brownlee. *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery, 2020.
- [10] William Wolberg. Breast Cancer Wisconsin (Original). UCI Machine Learning Repository, 1990. DOI: <https://doi.org/10.24432/C5HP4Z>.
- [11] Ronald T. Kneusel. *The art of randomness*. No Starch Press, San Francisco, 2024.